Updates in Rhea—a manually curated resource of biochemical reactions

Anne Morgat[1,2,*], Kristian B. Axelsen[1], Thierry Lombardot[1], Rafael Alcántara[3], Lucila Aimo[1], Mohamed Zerara[1], Anne Niknejad[4], Eugeni Belda[5], Nevila Hyka-Nouspikel[1], Elisabeth Coudert[1], Nicole Redaschi[1], Lydie Bougueleret[1], Christoph Steinbeck[3], Ioannis Xenarios[1,4,6] and Alan Bridge[1]

[1]Swiss-Prot Group, SIB Swiss Institute of Bioinformatics, Geneva, CH-1206, Switzerland, [2]Equipe BAMBOO, INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, F-38330, France, [3]Cheminformatics and Metabolism Team, European Bioinformatics Institute, Hinxton, CB10 1SD, UK, [4]Vital-IT, SIB Swiss Institute of Bioinformatics, Lausanne, CH-1015, Switzerland, [5]Genoscope—LABGeM, CEA, Evry, F-91057, France and [6]Department of Biochemistry, University of Geneva, Geneva, CH-1206, Switzerland

ABSTRACT

Rhea (http://www.ebi.ac.uk/rhea) is a comprehensive and non-redundant resource of expert-curated biochemical reactions described using species from the ChEBI (Chemical Entities of Biological Interest) ontology of small molecules. Rhea has been designed for the functional annotation of enzymes and the description of genome-scale metabolic networks, providing stoichiometrically balanced enzyme-catalyzed reactions (covering the IUBMB Enzyme Nomenclature list and additional reactions), transport reactions and spontaneously occurring reactions. Rhea reactions are extensively curated with links to source literature and are mapped to other publicly available enzyme and pathway databases such as Reactome, BioCyc, KEGG and UniPathway, through manual curation and computational methods. Here we describe developments in Rhea since our last report in the 2012 database issue of Nucleic Acids Research. These include significant growth in the number of Rhea reactions and the inclusion of reactions involving complex macromolecules such as proteins, nucleic acids and other polymers that lie outside the scope of ChEBI. Together these developments will significantly increase the utility of Rhea as a tool for the description, analysis and reconciliation of genome-scale metabolic models.

AIMS AND SCOPE OF RHEA

Rhea is a manually curated resource of biochemical reactions for the functional annotation of enzymes and the description of genome-scale metabolic networks (1). Rhea provides stoichiometrically balanced descriptions for enzyme-catalyzed reactions, transport reactions and spontaneously occurring reactions using chemical species from the Chemical Entities of Biological Interest (ChEBI) ontology (2), specifying reaction constituents, their stoichiometric coefficients and relative locations. This information is manually curated from peer-reviewed literature by experts. Each Rhea reaction is assigned a unique identifier, with uniqueness ensured by the calculation of a fingerprint for each reaction which considers the constituent compounds, their stoichiometry and localization. Reaction constituents are represented by the major micro-species at pH 7.3 (verified using the Marvin pKa calculator from ChemAxon (version 6.2.0, http://www.chemaxon.com)). All reactions are stoichiometrically balanced for both mass and charge, which facilitates the use of Rhea for the construction, analysis, comparison and reconciliation of genome-scale metabolic models (3,4). More details on the representation of reactions can be found in our preceding paper (1). Rhea provides metabolic reactions for a number of other biological data and knowledge resources including the EBI Enzyme Portal (5), the reference layer of the MetaboLights resource (6), the metabolic model analysis and reconciliation platform of MetaNetX.org (7,8), the microbial genomic annotation platform MicroScope (9) and IntEnz, a reference for the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecu-
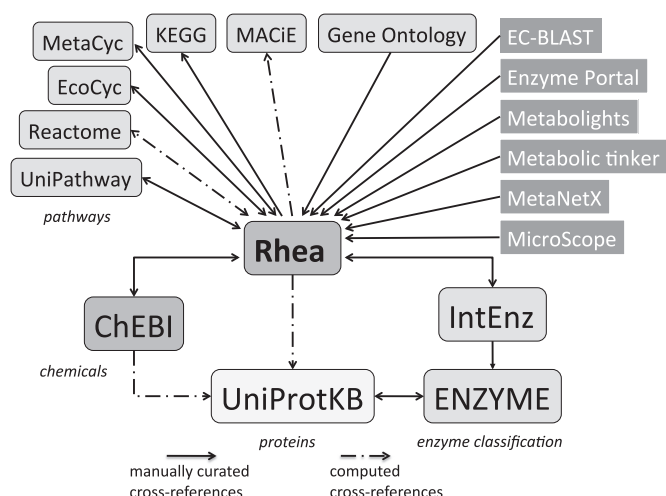
Figure 1. Interactions between Rhea and other resources. Rhea provides cross-references to chemical entities of ChEBI (2), to biochemical reactions of EcoCyc (19), MetaCyc (20), KEGG (18), MACiE (16), Reactome (17) and UniPathway (21), to EC numbers of IntEnz (10) and to protein sequences of UniProtKB (15). The Gene Ontology (GO) is closely aligned with ChEBI , and GO molecular functions describing enzymatic reactions cross-reference Rhea (29). Rhea is one of the reaction repositories employed by MicroScope (9), an integrated resource for the curation and comparative analysis of genomic and metabolic data of microbes. Rhea also provides metabolic reactions for a number of other resources including the EBI Enzyme Portal (5), the reference layer of the MetaboLights resource (6), the metabolic model analysis and reconciliation platform of MetaNetX.org (7,8), EC-BLAST (11) and Metabolic tinker (12).

lar Biology (NC-IUBMB) on the nomenclature and classification of enzymes (10). Rhea reactions are also used by tools such as EC-BLAST, a tool to automatically search and compare biochemical reactions (11) as well as Metabolic tinker (12), an online tool for guiding the design of synthetic metabolic pathways. Interactions between Rhea and other resources are described in Figure 1.
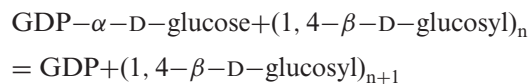
EXTENDING RHEA TO COMPLEX MACRO-MOLECULES AND POLYMERS

Computational models of cellular metabolism generally include instances of complex biological macromolecules such as proteins, nucleic acids and other polymers that lie outside the scope of ChEBI, which deals with small molecules and metabolites. To permit the representation of reactions involving such entities in Rhea, we have introduced generic compounds ('Rhea generics') and polymers ('Rhea polymers'). Rhea generics represent complex biological macromolecules such as proteins, nucleic acids and complex polysaccharides. Rhea polymers represent compounds that appear on both sides of a given reaction with different relative polymerization indices, such as 'n' and '$n + 1$'.

We describe first the use of Rhea generics. Each Rhea generic has a unique identifier and a name that specifies the nature of the biological macromolecule under consideration. Residues and functional groups that are modified during the course of the reaction are represented explicitly using entities from ChEBI, which allows stoichiometric balancing for mass and charge. An example

of the usage of Rhea generics is modification reactions involving acyl carrier protein (ACP), which plays an essential role in the process of fatty acid biosynthesis. Before ACP can accept acyl chains for elongation the protein must be activated by ACP synthase, which attaches a phosphopantetheine group from coenzyme A (CoA) to a conserved serine residue of ACP, releasing adenosine 3',5' bisphosphate (13). In Rhea, this post-translational modification is described by RHEA:12071 (Figure 2A). The substrate for this reaction is the Rhea generic 'apo-[ACP]' (GENERIC:9690), in which the target serine residue is represented by CHEBI:29999 ('L-serine residue'). The product of this reaction is holo-[ACP] (GENERIC:9685), which includes an O-(pantetheine-4′-phosphoryl)-L-serine residue represented by CHEBI:64479. The initiation of fatty acid synthesis on holo-[ACP] is represented by RHEA:41791, in which an acetyl group is transferred from acetyl CoA (CHEBI:57288) to the O-(pantetheine-4′-phosphoryl)-L-serine residue of the holo-[ACP] (GENERIC:9685), forming acetyl-[ACP] (GENERIC:9621) that includes an O-(S-acetylpantetheine-4′-phosphoryl)-L-serine residue (CHEBI:78446) (Figure 2B). Note that the same residue or group may appear in a number of distinct generic compounds and reactions (such as the L-serine residue CHEBI:29999). In some reactions several residues and/or functional groups from a single Rhea generic macromolecule may participate in the same chemical transformation. The yeast enzyme tRNA (cytidine^{32}/guanosine^{34}−2′-O)-methyltransferase (described in UniProtKB/Swiss-Prot record P38238) catalyzes the formation of 2′-O-methylribose at two sites in the anticodon loop of a single tRNA molecule (14). This reaction is RHEA:42399 (Figure 2C). Its substrate is Rhea GENERIC:10246, which includes both a cytidine 5′-phosphate residue (CHEBI:82748) at position 32 and a guanosine 5′-phosphate residue (CHEBI:74269) at position 34. The corresponding product is Rhea GENERIC:10247, which carries the corresponding 2′-O-methylcytidine 5′-phosphate residue (CHEBI:74495) and 2′-O-methylguanosine 5′-phosphate residue (CHEBI:74445) at the same positions.

Rhea polymers differ from Rhea generics. Rhea polymers have been introduced in order to allow balancing of polymerization reactions that include different abstract polymerization indices for polymers such as 'n' and '$n + 1$', as in this example:

$$GDP-\alpha-\text{D}-\text{glucose}+(1,4-\beta-\text{D}-\text{glucosyl})_n$$
$$= GDP+(1,4-\beta-\text{D}-\text{glucosyl})_{n+1}$$

ChEBI contains only a single Instance of each abstract polymer, with a single unknown polymerization index. Each Rhea polymer has an identifier (prefixed by 'POLYMER'), a name, a link to the corresponding ChEBI polymer and a relative polymerization index. Several Rhea polymers may share the same ChEBI entry, but they must have different polymerization indices, which are used in reaction balancing. The use of Rhea polymers in the context of polymerization reactions is shown in Figure 3.
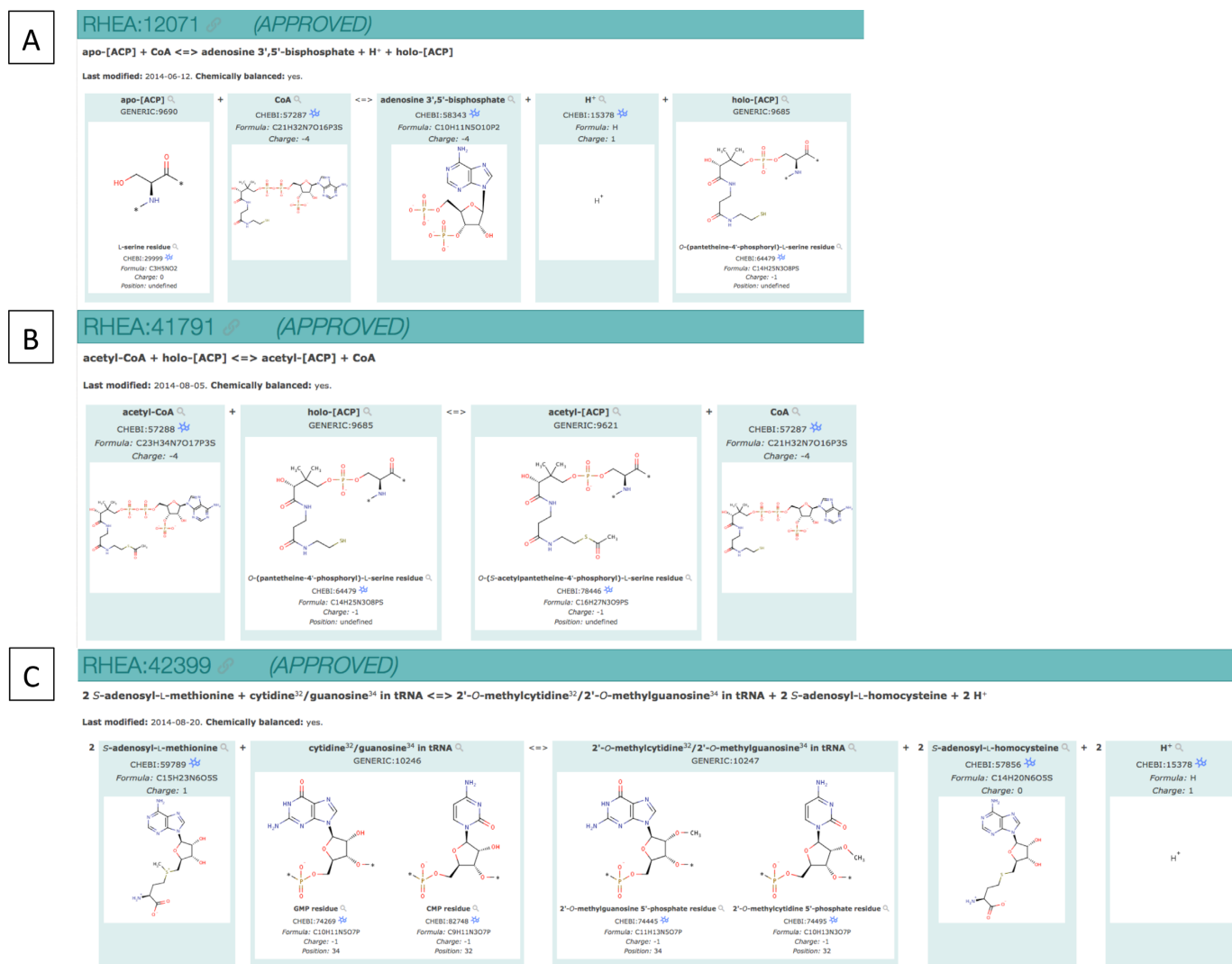
Figure 2. Rhea generics. (**A**) RHEA:12071 represents the priming of acyl carrier protein (ACP) by ACP synthase, where a phosphopantetheine group from coenzyme A (CoA) is attached to a conserved serine residue of ACP to form the activated holo-[ACP] and to release adenosine 3′,5′ bisphosphate (CHEBI:58343). Apo-[ACP] is modeled by a single L-serine residue (GENERIC:9690) and holo-[ACP] by an *O*-(pantetheine-4′-phosphoryl)-L-serine residue (GENERIC:9685). (**B**) RHEA:41791 represents the initiation of fatty acid synthesis on holo-[ACP] in which an acetyl group is transferred from acetyl CoA (CHEBI:57288) to the *O*-(pantetheine-4′-phosphoryl)-L-serine residue of the holo-[ACP] (GENERIC:9685), forming acetyl-[ACP] (GENERIC:9621) that includes an *O*-(*S*-acetylpantetheine-4′-phosphoryl)-L-serine residue (CHEBI:78446). (**C**) RHEA:42399 illustrates the use of Rhea generics composed of multiple residues. The substrate is Rhea GENERIC:10246, which includes both a CMP residue (CHEBI:82748) at position 32 and a GMP residue (CHEBI:74269) at position 34. The corresponding product is Rhea GENERIC:10247, which carries the corresponding 2′-*O*-methylcytidine 5′-phosphate residue (CHEBI:74495) and 2′-*O*-methylguanosine 5′-phosphate residue (CHEBI:74445) at the same positions.

RHEA CONTENT

Rhea has grown steadily since our last report through the expert curation of new chemical entities and reactions from peer-reviewed literature. At the time of writing, Rhea (release 53) includes 7044 unique reactions involving 5927 unique reaction participants, and cites 2766 unique PubMed identifiers (see http://www.ebi.ac.uk/rhea/statistics.xhtml for details). This corresponds to a 63% increase in the number of unique reactions and a 161% increase in the number of unique citations since our last publication in 2012 (Rhea release 24, containing 4321 unique reactions, 3788 unique reaction participants and citing 1058 unique PubMed identifiers).

The value and utility of Rhea reactions are enhanced by extensive cross-references to other public resources (Table 1). The cross-references are extensively manually curated and crosschecked, with information on possible corrections and clarifications being regularly exchanged between curators of Rhea and those of the other resources. In addition, cross-references are automatically added from Rhea to UniProtKB/Swiss-Prot (15) protein records (through Enzyme Commission (EC) numbers in IntEnz) and to reaction descriptions in MACiE (16) and Reactome (17) (through shared participants).
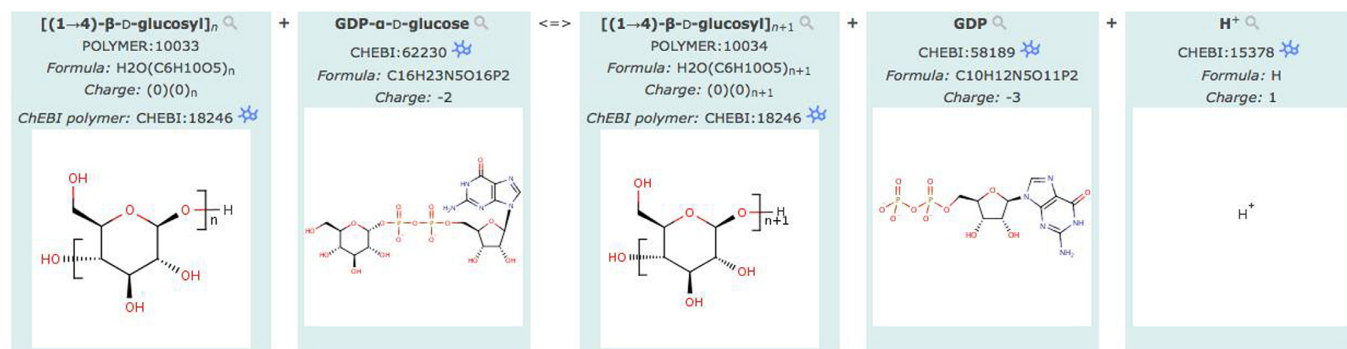
Figure 3. Rhea polymers. RHEA:17800 is a polymerization reaction where the glucosyl moiety of GDP-α-D-glucose (CHEBI:62230) is added to (1,4-β-D-glucosyl)$_n$ (POLYMER:10033) through a β-1$\rightarrow$4 linkage to produce (1,4-β-D-glucosyl)$_{n+1}$ (POLYMER:10034).

Table 1. Cross-references in Rhea (release 53)

Database	Ref.	Data type	Origin	Total	Unique
ChEBI http://www.ebi.ac.uk/chebi	(2)	Compound	Manually curated	36 928	6720
IntEnz http://www.ebi.ac.uk/intenz	(10)	EC number	Manually curated	4955	4110
MetaCyc http://metacyc.org	(20)	Reaction	Manually curated	4407	4279
EcoCyc http://ecocyc.org	(19)	Reaction	Manually curated	1136	1109
KEGG http://www.genome.jp/kegg	(18)	Reaction	Manually curated	3724	3641
UniPathway http://www.unipathway.org/pathway	(21)	Reaction	Manually curated	1878	1835
Reactome http://www.reactome.org	(17)	Reaction	Automatic via ChEBI IDs	608	582
UniProtKB www.uniprot.org	(15)	Protein	Automatic via EC numbers	22 0058	16 1289
PubMed http://www.ncbi.nlm.nih.gov/pubmed		Bibliographic citation	Manually curated	5164	2766

SUBMISSIONS TO RHEA

Rhea welcomes submissions describing new reactions or suggesting updates to existing reactions. All submissions should be posted on our SourceForge Reaction Requests/Updates tracker (http://sourceforge.net/p/rhea-ebi/reaction-requests-updates) with relevant information (name, 2D structure...) for each reaction participant and cross-references to other relevant databases and source literature where available. Reactions requested for a publication under review are assigned preliminary status during the peer-review process and acquire approved status once the manuscript has been accepted.

RHEA AVAILABILITY

The Rhea web server (http://www.ebi.ac.uk/rhea) provides programmatic access as well as browsing, searching and download facilities. Details of common search options — including compound names, compound and reaction identifiers, reaction equations, EC numbers, UniProtKB/Swiss-Prot accession numbers, bibliographic citations and identifiers from external cross-referenced resources such as KEGG (18), EcoCyc (19), MetaCyc (20), UniPathway (21), MACiE or Reactome — are provided in our last publication (1). Searches with compound identifiers may be pre-fixed with CHEBI, POLYMER or GENERIC to specify the desired type of molecule. Rhea generics and polymers may also be retrieved by searching for the associated ChEBI residue/group or compound (e.g. 'CHEBI:29999') or by name (e.g. 'ACP'). It is possible to link to reactions in the public web site using the following URL template http://www.ebi.ac.uk/rhea/reaction.xhtml?id=, adding the numerical reaction identifier as in this example: http://www.ebi.ac.uk/rhea/reaction.xhtml?id=10499.

All Rhea data is available for free download (http://www.ebi.ac.uk/rhea/download.xhtml) in BioPAX level 2 (biopax2) (22), RXN and RD (23) formats. In the BioPAX level 2 distribution of Rhea all reaction participants are defined by the class 'physicalEntityParticipant'. Cross-references to other databases such as ChEBI, EcoCyc, In-tEnz, KEGG, MACiE, MetaCyc, Reactome, UniPathway and UniProtKB are also available as tab-separated text files. The 2D structures of chemical compounds used in Rhea are available for download either as individual molfiles or as a Structure-Data File (SDF). These chemical formats are specified by Accelrys (formerly by Molecular Design Limited (MDL) (23)).

Rhea RESTful web services allow reactions to be retrieved in BioPAX level 2 (22), CMLReact (cmlreact) (24) or RXN CTfile (rxn) (23) formats by querying for their

identifier or other terms. Example queries are provided in the online documentation. Rhea also provides a BioJS component (BioJS.Rheaction, http://www.ebi.ac.uk/Tools/biojs/registry/Biojs.Rheaction.html), which can be used to display (and possibly modify the layout of) a Rhea reaction in an external web page given only its Rhea ID. The EBI Enzyme Portal makes use of this BioJS component to display reactions (example: http://www.ebi.ac.uk/enzymeportal/search/P45850/reactionsPathways).

FUTURE DIRECTIONS

We are actively developing Rhea as a vocabulary for the functional annotation of enzymes in UniProtKB. This annotation is currently provided using the enzyme classification (EC numbers) of the Enzyme Nomenclature committee of the IUBMB and textual reaction descriptions sourced from the ENZYME database (25) (itself derived from IntEnz). Our current work involves the translation of all outstanding IUBMB reactions (the majority of which involve generic compounds or polymers) into Rhea. We are also expanding Rhea to cover the hundreds of enzyme activities that are not yet described by the IUBMB classification (26–28), many of which already have textual reaction descriptions annotated in UniProtKB (one example being the aforementioned GDP-α-D-mannose hydrolysis reaction RHEA:28105 described in UniProtKB/Swiss-Prot record P32056). We will also exploit the underlying ontology of ChEBI in order to provide a logical reaction classification based on the curated relations between reaction participants. This will serve as a useful complement to the classification of enzymatic activities by the IUBMB.

REFERENCES

1. Alcantara,R., Axelsen,K.B., Morgat,A., Belda,E., Coudert,E., Bridge,A., Cao,H., de Matos,P., Ennis,M., Turner,S. *et al.* (2012) Rhea–a manually curated resource of biochemical reactions. *Nucleic Acids Res.*, **40**, D754–D760.
2. Hastings,J., de Matos,P., Dekker,A., Ennis,M., Harsha,B., Kale,N., Muthukrishnan,V., Owen,G., Turner,S., Williams,M. *et al.* (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, **41**, D456–D463.
3. El-Semman,I.E., Karlsson,F.H., Shoaie,S., Nookaew,I., Soliman,T.H. and Nielsen,J. (2014) Genome-scale metabolic reconstructions of *Bifidobacterium adolescentis* L2–32 and *Faecalibacterium prausnitzii* A2–165 and their interaction. *BMC Syst. Biol.*, **8**, 41.
4. Belda,E., Sekowska,A., Le Fevre,F., Morgat,A., Mornico,D., Ouzounis,C., Vallenet,D., Medigue,C. and Danchin,A. (2013) An updated metabolic view of the *Bacillus subtilis* 168 genome. *Microbiology*, **159**, 757–770.
5. Alcantara,R., Onwubiko,J., Cao,H., Matos,P., Cham,J.A., Jacobsen,J., Holliday,G.L., Fischer,J.D., Rahman,S.A., Jassal,B. *et al.* (2013) The EBI enzyme portal. *Nucleic Acids Res.*, **41**, D773–D780.
6. Haug,K., Salek,R.M., Conesa,P., Hastings,J., de Matos,P., Rijnbeek,M., Mahendraker,T., Williams,M., Neumann,S., Rocca-Serra,P. *et al.* (2013) MetaboLights–an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.*, **41**, D781–D786.
7. Ganter,M., Bernard,T., Moretti,S., Stelling,J. and Pagni,M. (2013) MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics*, **29**, 815–816.
8. Bernard,T., Bridge,A., Morgat,A., Moretti,S., Xenarios,I. and Pagni,M. (2014) Reconciliation of metabolites and biochemical reactions for metabolic networks. *Brief Bioinform*, **15**, 123–135.
9. Vallenet,D., Belda,E., Calteau,A., Cruveiller,S., Engelen,S., Lajus,A., Le Fevre,F., Longin,C., Mornico,D., Roche,D. *et al.* (2013) MicroScope–an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Res.*, **41**, D636–D647.
10. Fleischmann,A., Darsow,M., Degtyarenko,K., Fleischmann,W., Boyce,S., Axelsen,K.B., Bairoch,A., Schomburg,D., Tipton,K.F. and Apweiler,R. (2004) IntEnz, the integrated relational enzyme database. *Nucleic Acids Res.*, **32**, D434–D437.
11. Rahman,S.A., Cuesta,S.M., Furnham,N., Holliday,G.L. and Thornton,J.M. (2014) EC-BLAST: a tool to automatically search and compare enzyme reactions. *Nat. Methods*, **11**, 171–174.
12. McClymont,K. and Soyer,O.S. (2013) Metabolic tinker: an online tool for guiding the design of synthetic metabolic pathways. *Nucleic Acids Res.*, **41**, e113.
13. Elovson,J. and Vagelos,P.R. (1968) Acyl carrier protein. X. Acyl carrier protein synthetase. *J. Biol. Chem.*, **243**, 3603–3611.
14. Pintard,L., Lecointe,F., Bujnicki,J.M., Bonnerot,C., Grosjean,H. and Lapeyre,B. (2002) Trm7p catalyses the formation of two 2'-O-methylriboses in yeast tRNA anticodon loop. *EMBO J.*, **21**, 1811–1820.
15. UniProt Consortium. (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198.
16. Holliday,G.L., Andreini,C., Fischer,J.D., Rahman,S.A., Almonacid,D.E., Williams,S.T. and Pearson,W.R. (2012) MACiE: exploring the diversity of biochemical reactions. *Nucleic Acids Res.*, **40**, D783–D789.
17. Croft,D., Mundo,A.F., Haw,R., Milacic,M., Weiser,J., Wu,G., Caudy,M., Garapati,P., Gillespie,M., Kamdar,M.R. *et al.* (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **42**, D472–D477.
18. Kanehisa,M., Goto,S., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.
19. Keseler,I.M., Mackie,A., Peralta-Gil,M., Santos-Zavaleta,A., Gama-Castro,S., Bonavides-Martinez,C., Fulcher,C., Huerta,A.M., Kothari,A., Krummenacker,M. *et al.* (2013) EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.*, **41**, D605–D612.

20. Caspi,R., Altman,T., Billington,R., Dreher,K., Foerster,H., Fulcher,C.A., Holland,T.A., Keseler,I.M., Kothari,A., Kubo,A. *et al.* (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **42**, D459–D471.

21. Morgat,A., Coissac,E., Coudert,E., Axelsen,K.B., Keller,G., Bairoch,A., Bridge,A., Bougueleret,L., Xenarios,I. and Viari,A. (2012) UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res.*, **40**, D761–D769.

22. Demir,E., Cary,M.P., Paley,S., Fukuda,K., Lemer,C., Vastrik,I., Wu,G., D'Eustachio,P., Schaefer,C., Luciano,J. *et al.* (2010) The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.*, **28**, 935–942.

23. Dalby,A., Nourse,J.G., Hounshell,W.D., Gushurst,A.K.I., Grier,D.L., Leland,B.A. and Laufer,J. (1992) Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.*, **32**, 244–255.

24. Holliday,G.L., Murray-Rust,P. and Rzepa,H.S. (2006) Chemical markup, XML, and the world wide web. 6. CMLReact, an XML vocabulary for chemical reactions. *J. Chem. Inf. Model.*, **46**, 145–157.

25. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.

26. Ramkissoon,K.R., Miller,J.K., Ojha,S., Watson,D.S., Bomar,M.G., Galande,A.K. and Shearer,A.G. (2013) Rapid identification of sequences for orphan enzymes to power accurate protein annotation. *PLoS One*, **8**, e84508.

27. Shearer,A.G., Altman,T. and Rhee,C.D. (2014) Finding sequences for over 270 orphan enzymes. *PLoS One*, **9**, e97250.

28. Sorokina,M., Stam,M., Medigue,C., Lespinet,O. and Vallenet,D. (2014) Profiling the orphan enzymes. *Biol. Direct*, **9**, doi:10.1186/1745-6150-9-10.

29. Hill,D.P., Adams,N., Bada,M., Batchelor,C., Berardini,T.Z., Dietze,H., Drabkin,H.J., Ennis,M., Foulger,R.E., Harris,M.A. *et al.* (2013) Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology. *BMC Genomics*, **14**, 513.